Universidade Federal de Pernambuco

Graduação em Ciência da Computação
Centro de Informática

Utilizando Redes Neurais Artificiais para Predizer Erros em Inferência de Haplótipos

Proposta de Trabalho de Graduação

Aluno: Rafael Henrique da Silva Santos (rhss@cin.ufpe.br)
Orientador: Katia Silva Guimarães (katiag@cin.ufpe.br)

Recife, 19 de Setembro de 2011

1. Contexto

Inferência de Haplótipos (IH) é um grande desafio, pois ainda não foi proposto um modelo de sequenciador (*next-generation sequencing*) capaz de obter este tipo de dado diretamente no laboratório a custo razoável de tempo e recursos. Este tipo informação é valioso para o entendimento da evolução das espécies, assim como em estudos de associação, que visam correlacionar a ocorrência de certas doenças genéticas com padrões herdados nas células gaméticas (células haplóides). Diante da necessidade de se obter tal tipo de informação e da limitação dos sequenciadores hoje existentes, a solução encontrada é inferir combinatoriamente/estatisticamente haplótipos a partir de genótipos.

Muitos métodos para IH foram propostos [1][2][3][4], sendo aqueles que exploram certas estatísticas da população os que apresentaram melhor desempenho do ponto de vista da qualidade dos resultados. Infelizmente, as taxas de erros das abordagens ainda são muito altas, o que faz com que um viés significativo seja adicionado às análises que usam como base os haplótipos inferidos por tais modelos. Em um estudo prévio realizado por Rosa e Guimarães [5] verificou-se que, embora os algoritmos de IH apresentem taxas de erros próximas, diferentes abordagens tendem a errar em locais distintos das sequências de genótipos, uma vez que cada uma delas utiliza diferentes *insights* e estratégias para tentar resolver o problema. Reunir todas estas estratégias em um único software seria algo inviável tendo em vista a complexidade do problema e o fato de que muitos destes *insights* não podem ser aplicados concomitantemente, pois se contradizem em alguns aspectos.

Diante da problemática, surgiu a hipótese de que, se for possível caracterizar as regiões das sequências de genótipos em que cada método tende a errar, seria então possível desconsiderar as soluções de cada método para estas regiões, tornando possível um *ensemble* baseado em propriedades. Para isso, faz-se necessário estabelecer correlações entre as métricas de erro de inferência e as propriedades (características) dos genótipos.

2. Objetivo

O objetivo deste projeto é utilizar redes neurais artificiais para estabelecer correlações entre medidas de erro (como *Error Rate*, *Switch Error* e *Switch Distance*) de diversos métodos de IH e propriedades dos genótipos (como número de símbolos 2, nível de conservação, entre outras) e, assim, predizer aquelas. Como produto do projeto, será elaborado um programa para predizer erro em IH, onde o número de variáveis de erro e de propriedades será variável/parametrizável.

3. Cronograma

	Agosto		Setembro				Outubro				Novembro			
Atividade	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Definição do tema e levantamento bibliográfico														
Definição do escopo														
Construção das redes neurais														
Implementação do programa														
Elaboração do relatório														
Elaboração da apresentação														

4. Possíveis Avaliadores

- Aluizio Fausto Ribeiro Araújo

- Kátia Silva Guimarães

Referências

[1] Clark, A.: Inference of haplotypes from PCRamplified samples of diploid populations. Journal of Molecular Biology and Evolution 7, 111-122 (1990)

[2] Li, Z., Zhou, W., Zhang, X.S., Chen, L.: A parsimonious treegrow method for haplotype inference. Oxford Bioinformatics. 17, 3475–3481 (2005)

[3] Scheet, P., Stephens, M.: A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genótipos and haplotypic phase. American Journal of Human Genetics. 78, 629–644 (2006)

[4] Eronen, L., Geerts, F., Toivonen, H.: Haplorec: Efficient and accurate largescale reconstruction of haplotypes. BMC Bioinformatics. 7: 542 (2006)

[5] Rosa. R. S., Guimarães, K. S.: Insights on Haplotype Inference on Large Genotype Datasets. Lecture Notes in Bioinformatics. 6268, 47–58 (2010)

Assinaturas

_____

Kátia Silva Guimarães
Orientador

_____

Rafael Henrique da Silva Santos
Aluno